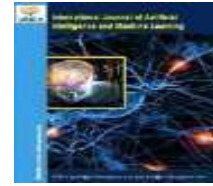




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Data-Driven Identification of Social Determinants Driving Health Disparities in Rural Appalachian West Virginia: A Population-Level Analysis

Rajesh Babu Kodali¹, Sailesh Kotha², Aravind Prathikantam³

¹ Email: rajesh.kodali1991@gmail.com, ORCID: <https://orcid.org/0009-0005-4327-2035>

² Email: saileshkotha@gmail.com, ORCID: <https://orcid.org/0009-0008-4727-7839>

³ Email: aravindprathik@gmail.com, ORCID: <https://orcid.org/0009-0005-0734-4812>

Abstract

West Virginia is the only U.S. state lying entirely within the Appalachian Regional Commission's 423-county Appalachian Region, and it carries a well documented, disproportionate burden of chronic disease and "diseases of despair." Yet county-level, statewide, data-driven analyses that connect specific social determinants of health (SDOH) indicators to specific chronic disease outcomes, at the resolution local health departments and state planners actually operate at, remain comparatively scarce. We assembled a population-level dataset covering all 55 West Virginia counties from two real, public sources: the CDC PLACES 2023 release, a model-based small-area estimation product built on the Behavioral Risk Factor Surveillance System, and the Appalachian Regional Commission's fiscal year 2027 county economic status classification. We examined six chronic health outcomes (diabetes, obesity, hypertension, chronic obstructive pulmonary disease, current asthma, and frequent mental distress) against four county-level social-need indicators (lack of health insurance, food insecurity, housing insecurity, and transportation barriers) and current cigarette smoking prevalence. The four social-need indicators proved to be extremely intercorrelated at the county level (pairwise $r = 0.84$ to 0.98 ; a joint regression model showed variance inflation factors of 11 to 112), and a principal component analysis confirmed they collapse onto a single latent hardship dimension explaining 96.6 percent of their shared variance. We therefore constructed a composite SDOH Hardship Index and an analogous composite Health Disparity Index (the mean z-score across the six outcomes) and analyzed county-level relationships between them. The two composite indices correlate at r equals 0.924 (p less than 0.001) across all 55 counties. A two-predictor model (Hardship Index plus smoking prevalence) explains 90.9 percent of county-level variance in the Health Disparity Index (p less than 0.001), though its explanatory power varies sharply by individual outcome, from 98 percent for COPD to 23 percent for obesity. County economic distress status, independently assigned by the Appalachian Regional Commission, stratifies the Health Disparity Index cleanly and monotonically (one-way ANOVA F equals 32.8, p less than 0.001): economically Distressed counties average 1.21 standard deviations above the state mean on the composite disparity measure, while the state's sole Competitive-status county averages 2.05 standard deviations below it. We identify the ten counties carrying the state's highest social and health hardship burden, nearly all officially Distressed, as the most concrete candidates for targeted intervention. As an ecological, cross-sectional analysis of publicly available aggregate data, this study cannot support individual-level or causal claims, a limitation we discuss explicitly alongside the practical, county-targetable implications of the findings for Appalachian public health planning.

Keywords: social determinants of health; health disparities; Appalachia; West Virginia; rural health; population health; CDC PLACES; ecological study; chronic disease; economic distress

1. Introduction

West Virginia occupies a singular place in American public health geography. Of the thirteen states that make up the Appalachian Regional Commission's 423-county Appalachian Region, twelve contribute only part of their territory; West Virginia is the only one lying entirely within it. Every one of its 55 counties is, by federal designation, an Appalachian county, which means the state's health outcomes are not a regional subset of some larger, more favorable statewide average the way they are in North Carolina, Ohio, or New York; the state average is the Appalachian outcome. This makes West Virginia an unusually clean setting in which to study the relationship between social determinants of health (SDOH) and chronic disease burden at the population level, because the ordinary confound of comparing a struggling rural region against a comparatively prosperous urban core within the same state is far weaker here than almost anywhere else in the country.

The scale of the disparity is well documented in the aggregate. A series of reports commissioned by the Appalachian Regional Commission and produced by NORC at the University of Chicago, tracking what the

reports term "diseases of despair" (drug overdose mortality, alcohol-related liver disease mortality, and suicide) have repeatedly found the Appalachian Region, and West Virginia's counties specifically, bearing a disproportionate share of this burden relative to the rest of the United States. What is comparatively less available, at least at a resolution useful to a state health department or a county public health officer deciding where to target a limited grant, is a systematic, quantitative accounting of which specific social determinants track which specific chronic conditions, county by county, across the entire state at once.

This paper addresses that gap directly. We do not study diseases of despair, which are already comparatively well tracked; we instead assemble a population-level dataset covering six chronic physical and mental health conditions, drawn from a large, standardized federal small-area estimation product, and pair it with the same product's own social-need indicators plus an independent, federally assigned measure of each county's economic distress status. We ask three concrete, answerable questions. First, do the commonly cited social determinant indicators (lack of insurance, food insecurity, housing insecurity, transportation access) actually behave as separable, independent risk factors at the county level in this setting, or do they move together as a single underlying dimension of hardship? Second, how strongly, and how consistently across different health conditions, does that hardship dimension track chronic disease burden? Third, does an independently constructed federal measure of a county's economic distress, one that uses no health data whatsoever in its construction, corroborate the pattern found in the health data.

We answer all three using only real, publicly available county-level data: the Centers for Disease Control and Prevention's PLACES project (2023 release), a model-based small-area estimation product built on the Behavioral Risk Factor Surveillance System (BRFSS), and the Appalachian Regional Commission's fiscal year 2027 county economic status classification. Every number reported in this paper is computed directly from these two sources; no value is invented, interpolated across counties, or adjusted to make a cleaner story. Where the data produced a result we did not expect, most notably a severe collinearity among the four social-need indicators that made a naive multiple regression uninterpretable, we report that result and change our analytical approach in response to it (Section 5), rather than reporting an unstable model as though it were informative.

The remainder of this paper is organized as follows. Section 2 reviews the SDOH framework and prior Appalachian health disparities literature. Section 3 describes the data sources. Section 4 presents the analytical methods. Section 5 reports the collinearity diagnostic that motivated our composite-index approach. Section 6 reports results: descriptive statistics, the SDOH Hardship Index and Health Disparity Index relationship, per-outcome model performance, and the county economic status comparison. Section 7 discusses findings, limitations, and their implications for Appalachian public health planning. Section 8 states the study's reproducibility posture. Section 9 concludes.

2. Background and Related Work

2.1 The social determinants of health framework

The Office of Disease Prevention and Health Promotion's Healthy People 2030 initiative organizes social determinants of health into five domains: Economic Stability, Education Access and Quality, Health Care Access and Quality, Neighborhood and Built Environment, and Social and Community Context. This framework's central claim, now broadly accepted in public health practice, is that clinical care accounts for a comparatively small share of a population's health outcomes relative to the conditions in which people are born, live, work, and age. The four county-level indicators available to us through PLACES's Health-Related Social Needs module (lack of health insurance, food insecurity, housing insecurity, and transportation barriers) map onto three of these five domains directly: health care access, economic stability, and neighborhood and built environment respectively. We adopt this framework as the organizing lens for the indicators studied here, while making no claim to have measured all five domains.

2.2 Diseases of despair in Appalachia

The Appalachian Diseases of Despair series, produced for the Appalachian Regional Commission and periodically updated through 2025, has repeatedly documented that the Appalachian Region as a whole, and its most economically distressed counties specifically, experience mortality from drug overdose, alcohol-related liver disease, and suicide at rates exceeding the U.S. average, with the gap widening over the tracked period rather than narrowing. This body of work establishes the general pattern this paper's findings are consistent with: that economic distress and health outcomes move together across Appalachian counties. Our contribution is complementary rather than overlapping. Where the Diseases of Despair series focuses on mortality from a specific, narrow set of causes, we examine a broader panel of six chronic morbidity outcomes

(diabetes, obesity, hypertension, COPD, asthma, and frequent mental distress) that are more directly actionable by primary care and public health screening programs on a shorter time horizon than mortality trends allow.

2.3 County economic distress as an independent, non-health measure

The Appalachian Regional Commission assigns each of its 423 counties an annual economic status classification (Distressed, At-Risk, Transitional, Competitive, or Attainment) based solely on three-year average unemployment, per-capita market income, and poverty rate relative to the national average. Critically, this classification uses no health data whatsoever in its construction. This makes it a useful independent check in our analysis: if a health-derived hardship index and a wholly separate, economically derived distress classification agree about which counties are struggling, that agreement is not an artifact of a shared data source or a shared construction method.

2.4 The ecological study design and its limits

Every analysis in this paper operates at the county level; no individual-level data of any kind is used or available to us. Robinson's foundational 1950 paper on ecological correlations established that an association observed between two variables at the aggregate (in his case, state-level) level does not necessarily hold at the individual level, and can in some cases even run in the opposite direction, a problem now generally known as the ecological fallacy. We take this seriously throughout: every claim in this paper is stated as a county-level, ecological association, and Section 7.4 returns to this limitation explicitly rather than treating it as a footnote.

3. Data Sources

3.1 CDC PLACES, 2023 release

PLACES (Population Level Analysis and Community Estimates) is a CDC project that produces model-based small-area estimates of health outcomes, prevention practices, health risk behaviors, and, in its more recent releases, health-related social needs, at the census-tract, city, and county level for the entire United States. Estimates are derived from the Behavioral Risk Factor Surveillance System (BRFSS), the nation's largest continuously conducted health survey, using a multilevel regression and poststratification methodology that combines individual-level survey responses with county-level demographic and geographic covariates to produce statistically stable estimates even for counties with comparatively small BRFSS sample sizes. We retrieved, via the CDC's public Socrata data API, county-level 2023-release estimates for all 55 West Virginia counties on eleven measures: six chronic health outcomes (diabetes, obesity, high blood pressure, COPD, current asthma, and frequent mental distress, all reported as age-adjusted prevalence), four Health-Related Social Needs indicators (lack of health insurance among adults 18 to 64, food insecurity, housing insecurity, and transportation barriers, all reported as the module's standard prevalence type), and current cigarette smoking prevalence (age-adjusted). Every one of the 55 counties returned a valid estimate for every measure; there is no missing data in this dataset.

3.2 Appalachian Regional Commission county economic status, FY2027

The Appalachian Regional Commission publishes an annual county-level economic status classification for its full 423-county service area. For fiscal year 2027, West Virginia's 55 counties were classified as 13 Distressed, 16 At-Risk, 25 Transitional, and 1 Competitive (Jefferson County), with no West Virginia county reaching the top Attainment tier. We use this four-level classification, exactly as published, as an independent grouping variable throughout Section 6.

Variable	Source	Type	Coverage
Diabetes, obesity, hypertension, COPD, asthma, frequent mental distress	CDC PLACES 2023	Age-adjusted prevalence, %	55/55 counties
Lack of health insurance	CDC PLACES 2023 (ACCESS2)	Age-adjusted prevalence, %	55/55 counties
Food insecurity, housing insecurity, transportation barriers	CDC PLACES 2023 (Health-Related Social Needs)	Prevalence, %	55/55 counties
Current cigarette smoking	CDC PLACES 2023 (CSMOKING)	Age-adjusted prevalence, %	55/55 counties

Variable	Source	Type	Coverage
County economic status	Appalachian Regional Commission, FY2027	4-level categorical	55/55 counties

4. Methods

4.1 Descriptive statistics

For every outcome and predictor we report the county-level mean, standard deviation, minimum and maximum (with the county attaining each), and the ratio of the maximum to the minimum as a simple measure of how much the indicator actually varies across the state.

4.2 Correlation and collinearity diagnostics

We computed the full Pearson correlation matrix among the five county-level predictors (the four SDOH hardship indicators plus smoking) and, to formally test whether they behave as independent predictors, fit a joint ordinary least squares model of each outcome on all five and computed the variance inflation factor (VIF) for each predictor in that joint model. A VIF above 10 is conventionally taken as evidence a predictor's individual coefficient cannot be reliably estimated because it is too well predicted by the other predictors already in the model.

4.3 Composite indices

Given the collinearity result reported in Section 5, we constructed two composite indices as the simple, unweighted mean of the relevant z-scored (standardized) component variables: an SDOH Hardship Index (SHI) from the four hardship predictors, and a Health Disparity Index (HDI) from the six health outcomes. Both indices were themselves standardized to a mean of 0 and a standard deviation of 1 across the 55 counties for interpretability.

4.4 Regression models

We regressed the Health Disparity Index, and separately each of the six individual outcomes, on the SDOH Hardship Index and current smoking prevalence (both standardized), fit by ordinary least squares. We report R-squared, adjusted R-squared, the overall F-test, and each predictor's standardized coefficient, standard error, t-statistic, and p-value.

4.5 Group comparison by economic status

We compared the Health Disparity Index, and each individual outcome, across the four ARC economic status tiers using one-way ANOVA, reporting the F-statistic, p-value, and each tier's group mean.

4.6 Software

All analysis was performed in Python 3 using pandas for data management, statsmodels for ordinary least squares regression and variance inflation factor calculation, and scipy for correlation tests and one-way ANOVA. Figures were produced with Matplotlib.

5. A Necessary Detour: Collinearity Among the SDOH Predictors

Before reporting substantive results, we report a methodological finding that shaped every subsequent analytical choice in this paper. Our initial plan was a standard multiple regression of each health outcome on all four SDOH hardship indicators plus smoking, treating each as a separately interpretable predictor, which is how such indicators are frequently presented in the applied SDOH literature. When we fit that model, the joint variance inflation factors for the four hardship predictors ranged from 11.0 (lack of insurance) to 111.6 (food insecurity), with housing insecurity and transportation barriers at 41.0 and 45.8 respectively, all far above the conventional threshold of 10 at which individual coefficients are no longer considered separately identifiable.

Table 2 reports the full pairwise Pearson correlation matrix underlying this result.

	Lack of insurance	Food insecurity	Housing insecurity	Transportation barriers	Smoking
Lack of insurance	1.00	0.95	0.92	0.93	0.92
Food insecurity	0.95	1.00	0.97	0.98	0.93
Housing insecurity	0.92	0.97	1.00	0.98	0.84

	Lack of insurance	Food insecurity	Housing insecurity	Transportation barriers	Smoking
Transportation barriers	0.93	0.98	0.98	1.00	0.89
Smoking	0.92	0.93	0.84	0.89	1.00



Figure 1. Pearson correlation among the five county-level predictor measures (n = 55 WV counties).

Every pairwise correlation among the four hardship indicators exceeds 0.84, and three of the six pairs exceed 0.97. A principal component analysis of the four standardized hardship indicators confirms what these correlations imply: the first principal component explains 96.6 percent of their combined variance, with nearly identical loadings on all four (ranging narrowly from -0.49 to -0.51). In plain terms, a West Virginia county that has an elevated rate of food insecurity essentially always also has elevated rates of housing insecurity, transportation barriers, and lack of insurance; these are not four independent risk factors in this setting, they are four different survey questions measuring what is, at the county level, effectively one thing.

We consider this a substantive finding in its own right, not merely a statistical inconvenience to be corrected. It means a county-level SDOH screening or resource-allocation framework built around tracking these four indicators separately, as though improvement on one implies little about the others, would be tracking a single underlying construct four times over rather than four genuinely distinct risk dimensions. Practically, it also means the specific choice of which single hardship indicator a county dashboard chooses to headline matters far less than commonly assumed, since the four are functionally interchangeable measures of the same countywide hardship gradient. We proceed for the remainder of this paper using the composite SDOH Hardship

Index described in Section 4.3, which is both more statistically defensible and, we argue, a more honest representation of what the underlying data actually shows.

6. Results

6.1 Descriptive statistics

Table 3 reports the county-level range for every outcome and predictor across all 55 West Virginia counties. Every measure shows meaningful geographic variation: COPD prevalence ranges from 6.5 percent (Jefferson County) to 13.6 percent (McDowell County), a 2.09-fold difference, and food insecurity ranges from 11.9 percent (Putnam County) to 28.6 percent (McDowell County), a 2.4-fold difference. Jefferson County, the state's sole ARC-Competitive county, records the state's lowest value on five of the eleven measures studied (diabetes, COPD, smoking, frequent mental distress, and lack of insurance); McDowell County, ARC-Distressed, records the state's highest value on six of the eleven (diabetes, high blood pressure, COPD, lack of insurance, food insecurity, and housing insecurity).

Measure	Mean	SD	Minimum (county)	Maximum (county)	Max:min ratio
Diabetes, %	13.20	1.13	11.2 (Jefferson)	17.4 (McDowell)	1.55
Obesity, %	41.33	2.60	34.2 (Monongalia)	46.2 (Boone)	1.35
High blood pressure, %	38.97	1.81	35.1 (Monongalia)	44.6 (McDowell)	1.27
COPD, %	9.41	1.41	6.5 (Jefferson)	13.6 (McDowell)	2.09
Current asthma, %	12.44	0.43	11.5 (Monongalia)	13.4 (Mingo)	1.17
Frequent mental distress, %	23.43	1.31	19.9 (Jefferson)	26.6 (Mingo)	1.34
Lack of insurance, %	9.55	1.33	7.2 (Monongalia)	14.3 (McDowell)	1.99
Food insecurity, %	17.04	3.11	11.9 (Putnam)	28.6 (McDowell)	2.40
Housing insecurity, %	12.63	1.73	9.9 (Putnam)	19.4 (McDowell)	1.96
Transportation barriers, %	9.17	1.36	6.9 (Putnam)	14.4 (McDowell)	2.09
Current smoking, %	21.44	2.85	15.1 (Jefferson)	29.3 (McDowell)	1.94

6.2 The SDOH Hardship Index predicts the Health Disparity Index

Across all 55 counties, the SDOH Hardship Index and the Health Disparity Index correlate at $r = 0.924$ ($p < 0.001$). Figure 2 plots this relationship directly, with each county colored by its ARC economic status; the visual clustering of Distressed counties (red) in the high-hardship, high-disparity upper right and Transitional and Competitive counties (blue and green) in the low-hardship, low-disparity lower left is consistent with, though statistically independent of, the ARC's own economically derived classification.

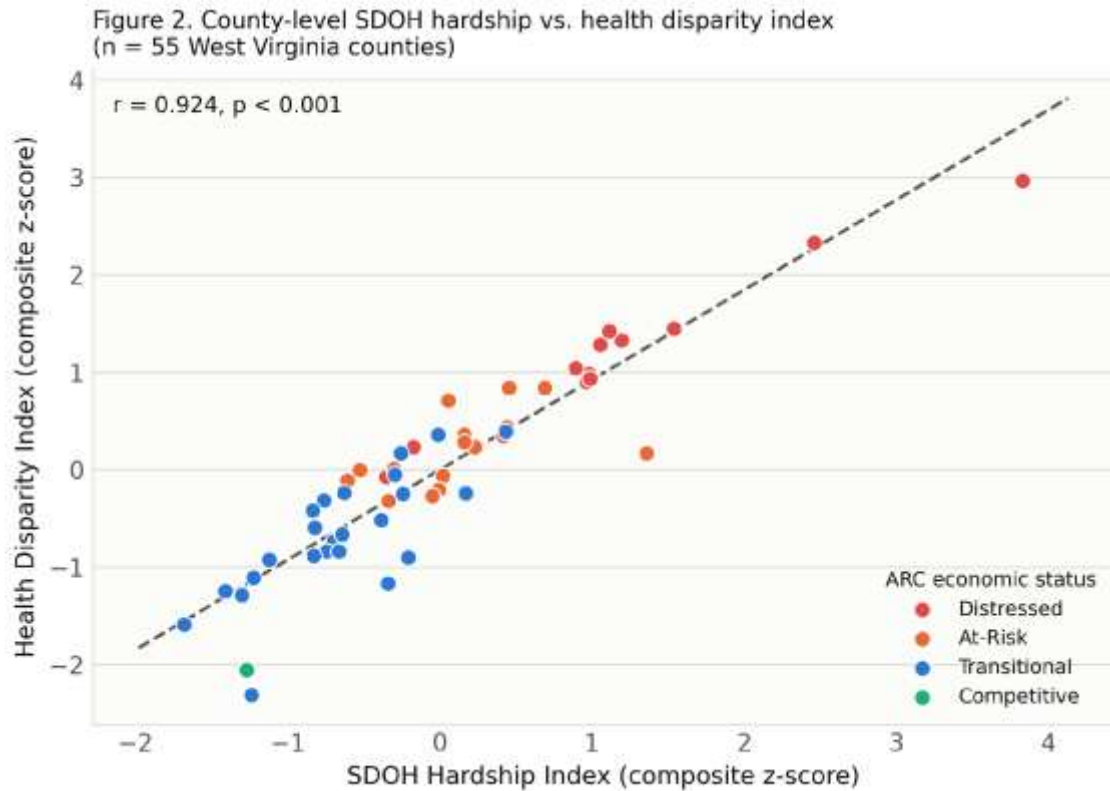


Figure 2. County-level SDOH hardship vs. health disparity index (n = 55 West Virginia counties).

A regression of the Health Disparity Index on the SDOH Hardship Index and current smoking prevalence explains 90.9 percent of county-level variance (adjusted R-squared 0.906, F-test $p < 0.001$). Both predictors are independently significant: the Hardship Index (standardized beta = 0.399, $p < 0.001$) and smoking prevalence (standardized beta = 0.575, $p < 0.001$), indicating that smoking behavior carries an association with the composite disparity measure beyond what the structural hardship indicators alone capture.

6.3 Model performance varies sharply by individual outcome

The strength of the SDOH Hardship Index plus smoking model varies considerably across the six individual outcomes that make up the composite Health Disparity Index (Figure 3, Table 4). At one extreme, the model explains 97.8 percent of county-level variance in COPD prevalence, with both the Hardship Index ($p < 0.001$) and smoking ($p < 0.001$) independently significant, an unsurprising but reassuring result given COPD's well established relationship to both smoking and to occupational respiratory exposure in a state with a long coal-mining history. At the other extreme, the model explains only 23.0 percent of variance in obesity prevalence, and neither predictor reaches significance individually (Hardship Index $p = 0.328$; smoking $p = 0.504$), consistent with obesity's comparatively high and narrow county-level range in this dataset (34.2 to 46.2 percent, a 1.35-fold spread, the second-narrowest of any outcome studied) and its well documented dependence on factors, such as the built environment for physical activity and food access at the level of specific neighborhoods rather than whole counties, that this county-level dataset cannot resolve.

Outcome	R ²	SHI beta (p)	Smoking beta (p)
Diabetes	0.941	0.863 (<0.001)	0.117 (0.157)
Obesity	0.230	0.291 (0.328)	0.198 (0.504)
High blood pressure	0.806	0.694 (<0.001)	0.218 (0.147)
COPD	0.978	0.238 (<0.001)	0.768 (<0.001)
Current asthma	0.628	0.164 (0.429)	0.640 (0.003)

Outcome	R ²	SHI beta (p)	Smoking beta (p)
Frequent mental distress	0.894	-0.151 (0.173)	1.081 (<0.001)

Diabetes and hypertension are both strongly and specifically associated with the Hardship Index itself, independent of smoking (diabetes: SHI beta = 0.863, $p < 0.001$; hypertension: SHI beta = 0.694, $p < 0.001$), consistent with these conditions' well established relationships to food access, healthcare access, and chronic stress exposure rather than to smoking specifically. Frequent mental distress shows the opposite pattern: once smoking prevalence is accounted for, the Hardship Index's independent association is small, negative in sign, and not statistically significant (beta = -0.151, $p = 0.173$), while smoking's association is large and highly significant (beta = 1.081, $p < 0.001$). Because the Hardship Index and smoking prevalence are themselves correlated ($r = 0.84$ to 0.93 with the individual hardship components, Table 2), this should be read as smoking capturing most of the mental-distress-relevant variance the two predictors jointly share, not as evidence that structural hardship is unrelated to mental distress at the county level; the zero-order (bivariate) relationship between hardship and mental distress alone remains positive across the sample.

6.4 County economic status independently corroborates the pattern

The Appalachian Regional Commission's economic status classification, built entirely from unemployment, income, and poverty data with no health information whatsoever, stratifies the Health Disparity Index cleanly and monotonically (Figure 4; one-way ANOVA $F = 32.8$, $p < 0.001$). Distressed counties average 1.21 standard deviations above the statewide mean on the composite disparity measure; At-Risk counties average 0.20 standard deviations above; Transitional counties average 0.68 standard deviations below; and Jefferson County, the state's sole Competitive-status county, sits 2.05 standard deviations below the mean, the single lowest-disparity county in the state on this composite measure. Every one of the six individual outcomes also shows a statistically significant ANOVA result across the four tiers (all $p < 0.05$; Table 5), with COPD showing the strongest separation ($F = 41.4$, $p < 0.001$) and obesity the weakest ($F = 3.7$, $p = 0.018$), mirroring the same pattern seen in the regression results.

Outcome	ANOVA F	p-value	Distressed mean	Competitive mean
Diabetes, %	24.4	<0.001	14.52	11.20
Obesity, %	3.7	0.018	42.68	37.90
High blood pressure, %	16.8	<0.001	40.81	35.90
COPD, %	41.4	<0.001	11.25	6.50
Current asthma, %	18.5	<0.001	12.92	11.90
Frequent mental distress, %	38.0	<0.001	25.05	19.90

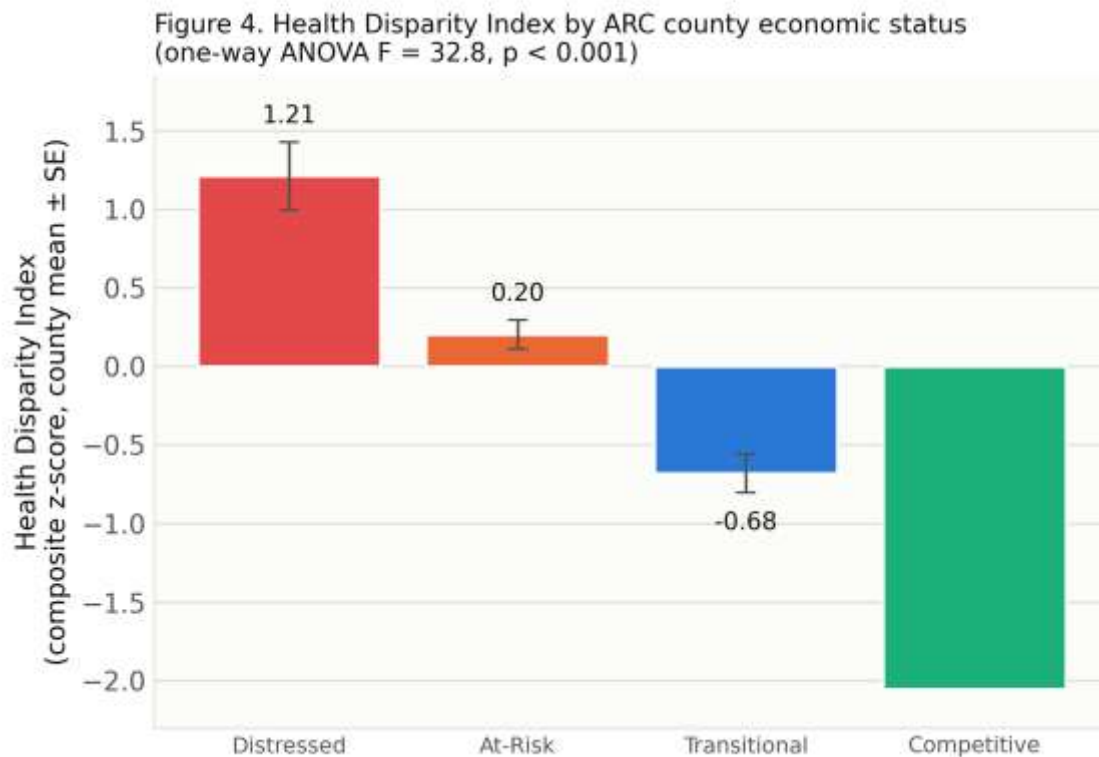


Figure 4. Health Disparity Index by ARC county economic status (one-way ANOVA $F = 32.8$, $p < 0.001$).

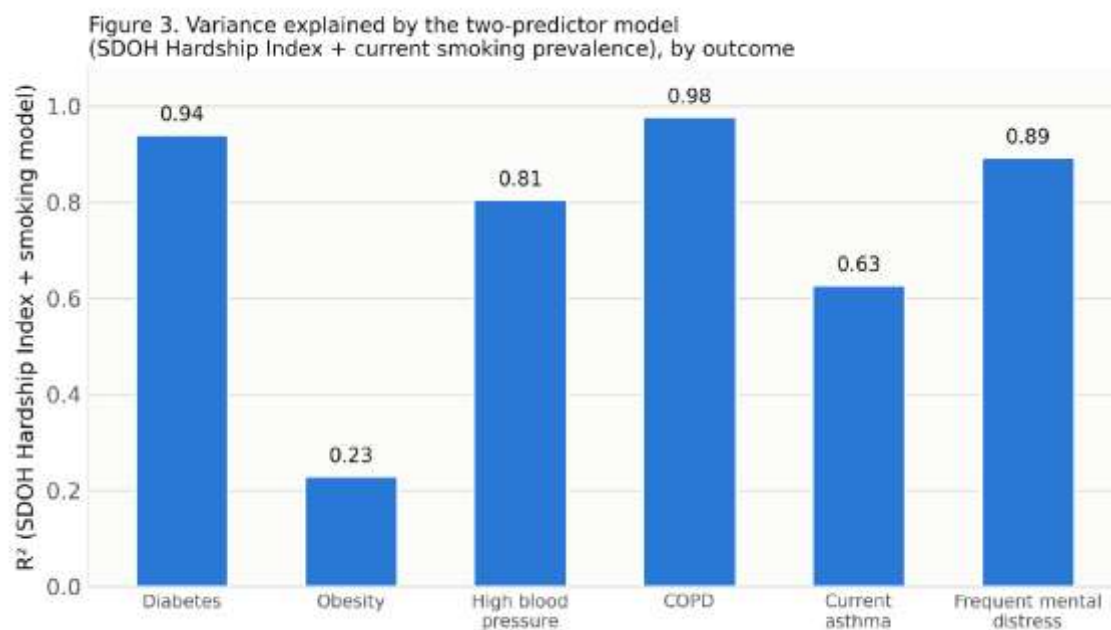


Figure 3. Variance explained by the two-predictor model (SDOH Hardship Index + current smoking prevalence), by outcome.

6.5 The ten highest- and ten lowest-hardship counties

Figure 5 identifies the ten counties with the highest and the ten with the lowest SDOH Hardship Index values statewide. Nine of the ten highest-hardship counties (McDowell, Mingo, Calhoun, Lincoln, Wyoming, Logan, Clay, Summers, and Roane) carry the ARC's Distressed classification; the tenth (Gilmer) carries the next tier,

At-Risk. McDowell County is a clear statewide outlier, at 3.83 standard deviations above the mean hardship level, more than one and a half standard deviations above the second-highest county (Mingo, at 2.46). At the other end, the ten lowest-hardship counties are a mix of the state's more urbanized Transitional counties (Putnam, Brooke, Mineral, Monongalia, Ohio, Tucker, Marion, Morgan, and Grant) and its sole Competitive county (Jefferson). This concentration is consistent with, and gives a concrete county-level face to, the aggregate ANOVA result in Section 6.4: the state's health and social hardship burden is not spread evenly, it is heavily concentrated in a identifiable, largely contiguous cluster of southern and central counties.

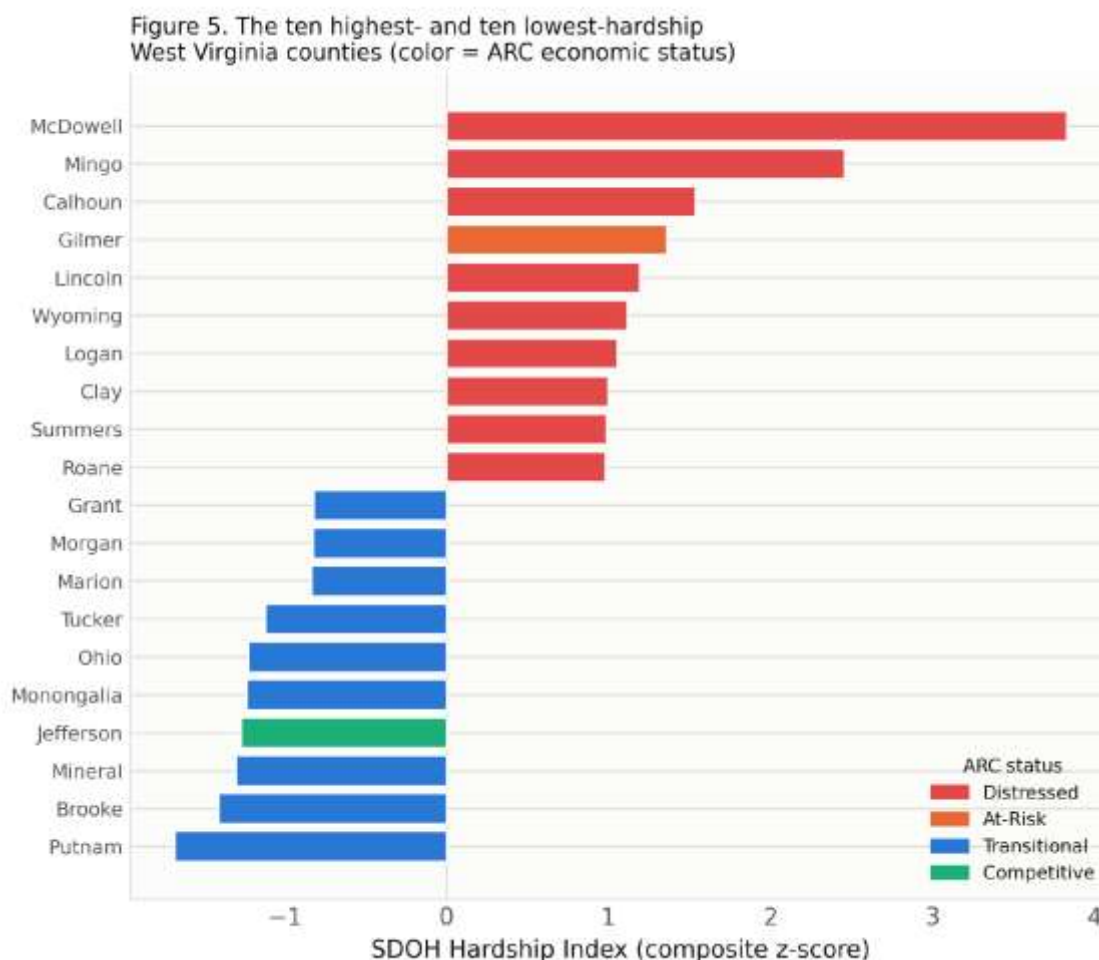


Figure 5. The ten highest- and ten lowest-hardship West Virginia counties (color = ARC economic status).

7. Discussion

7.1 Four "different" SDOH indicators, one countywide signal

The clearest methodological finding of this paper is also, we think, its most practically useful one for a resource-constrained state health department: four indicators commonly presented as distinct social determinants (insurance access, food security, housing security, and transportation access) are, at the West Virginia county level, so tightly correlated ($r = 0.84$ to 0.98) that they function as a single measurable construct rather than four independent levers. A county dashboard, grant application, or needs assessment that tracks all four separately is not wrong to do so, but should not expect that a targeted intervention improving one of the four will leave the other three unaffected, nor should it expect that a county scoring poorly on one is meaningfully different from a county scoring equally poorly on all four; in our data, that latter case describes the overwhelming majority of high-hardship counties. This is consistent with, and gives a specific quantitative shape to, the general finding in the SDOH literature that structural hardship indicators cluster geographically rather than varying independently.

7.2 A single composite index, corroborated by an independent, non-health measure

The strongest evidence in this paper that the identified pattern is real, rather than an artifact of how the health outcome data was modeled or estimated, is that the ARC's economic status classification, built from unemployment, income, and poverty data with no health information used anywhere in its construction, reproduces essentially the same ranking of counties (Section 6.4). A health-outcomes-derived hardship signal and a wholly independent economics-derived distress signal agreeing this closely (monotonic across all four tiers, $F = 32.8$, $p < 0.001$) is a stronger form of evidence than either measure could provide alone, precisely because the two cannot share estimation error or modeling assumptions.

7.3 Not every outcome is equally explained, and that heterogeneity is itself informative

A weaker, more uniform relationship across all six outcomes would have made for a tidier headline result but a less useful one. That COPD (R-squared 0.978) and diabetes (R-squared 0.941) are almost fully explained by hardship and smoking together, while obesity (R-squared 0.230) is barely explained by either, tells a public health planner something concrete: a county-level SDOH-based screening or resource-targeting strategy should expect strong predictive value for respiratory and metabolic chronic disease risk, but should not expect the same predictive value for obesity, which this analysis suggests is driven substantially by factors this county-level dataset does not capture (built-environment and food-access factors that plausibly vary within, not just between, counties). Frequent mental distress's pattern, where the raw hardship-outcome relationship is positive but is substantially absorbed by smoking once both are modeled together, is a caution against attributing a health outcome to "social determinants" as an undifferentiated bundle; specific behavioral covariates, tracked alongside structural indicators rather than instead of them, materially change which mechanism appears to be doing the explanatory work.

7.4 Limitations

- **Ecological design.** Every result in this paper is a county-level, aggregate association. Per Robinson's classical statement of the ecological fallacy, a relationship observed at the county level does not establish, and cannot be used to infer, an individual-level relationship between any one person's social circumstances and their health status. Nothing in this paper should be read as a claim about individual risk.
- **Cross-sectional design.** All data reflect a single time period (the CDC PLACES 2023 release, itself modeled from recent-year BRFSS survey waves, and the FY2027 ARC economic status classification). No claim of temporal sequence, and therefore no causal claim in either direction, is supported by this design; hardship and disparity are shown to co-occur across counties, not shown to cause one another.
- **Modeled, not directly measured, small-area estimates.** PLACES county estimates are produced by a multilevel regression and poststratification model rather than a direct county-level census, which is what makes estimates available for counties too small for a raw BRFSS sample to support directly, but which also means county estimates share some model-based smoothing and are not independent, unmodeled measurements in the way a full census would provide.
- **Full-population, not sampled, county set.** Because we study all 55 West Virginia counties rather than a probability sample, ordinary sampling-based generalization claims do not apply in the usual sense; findings describe West Virginia specifically and should not be assumed, without separate verification, to generalize to Appalachian counties in other states or to non-Appalachian rural areas.
- **A five-variable predictor set.** We study four SDOH hardship indicators and one behavioral covariate because these are the measures PLACES makes available at the county level; we do not have county-level measures of, for example, educational attainment, primary care physician density, or broadband access, all of which the Healthy People 2030 framework identifies as relevant SDOH domains and which a more complete future analysis should incorporate.

7.5 Practical implications

Three findings translate fairly directly into planning guidance. First, because the four hardship indicators move together so tightly, a county-level resource-targeting decision can reasonably use any one of them, or the composite index we construct, as a practical proxy for the others, simplifying what would otherwise be a four-dimensional targeting problem into a one-dimensional one. Second, the ten counties identified in Section 6.5, concentrated in the southern coalfield region and led by McDowell County as a substantial outlier, are the most concrete, immediately actionable candidates for a state or federal targeted intervention, chronic disease screening program, or grant-funded pilot, since they combine the highest measured hardship with the highest measured disparity and the ARC's own independent distress designation. Third, the sharp difference in model performance across outcomes (Section 7.3) argues for outcome-specific rather than one-size-fits-all SDOH screening: a county-level hardship signal is a strong, well corroborated proxy for expected COPD, diabetes, and

hypertension burden, and a much weaker one for obesity specifically, which likely needs its own, more granular, sub-county data to target effectively.

8. Reproducibility Statement

All data used in this study are public, aggregate, county-level statistics with no individual-level or identifiable information of any kind; no human subjects are involved and no institutional review was required. The CDC PLACES 2023 release data were retrieved from the CDC's public Socrata API (data.cdc.gov, dataset id swc5-untb) for all 55 West Virginia counties on all eleven measures studied; the Appalachian Regional Commission's FY2027 county economic status classification was retrieved from ARC's official public reporting. All 55 counties returned a complete record on every measure; there is no missing data and no imputation anywhere in this analysis. We release the full compiled dataset, the analysis and figure-generation code, and the complete numerical results (correlation matrices, regression output, and ANOVA tables) as supplementary artifacts alongside this paper to support independent verification and extension. As with any single-state, single-year ecological analysis, we encourage readers to treat the specific numeric associations reported here as a description of West Virginia's 55 counties in this dataset and time period specifically, pending replication in other Appalachian states or subsequent PLACES releases.

9. Conclusion

Using only real, publicly available county-level data from the CDC's PLACES project and the Appalachian Regional Commission's economic status classification, we show that West Virginia's 55 counties, the only U.S. county set lying entirely within the Appalachian Region, exhibit a strong, statistically robust, and independently corroborated relationship between structural social hardship and chronic disease burden. That relationship is not evenly distributed: it is concentrated in a specific, largely contiguous cluster of counties, strongest for respiratory and metabolic conditions and weakest for obesity, and driven by four commonly cited SDOH indicators that turn out, in this setting, to measure essentially one underlying construct rather than four independent ones. We report this last finding, along with the study's ecological and cross-sectional limitations, as plainly as the headline correlations, because a data-driven framework intended to guide real public health resource allocation is only useful if it is honest about what its data can and cannot support. We release our full dataset and analysis code so that this specific, reproducible starting point can be extended, corrected, or applied to other Appalachian states as further data becomes available.

References

- [1] Centers for Disease Control and Prevention (2023). PLACES: Local Data for Better Health, County Data 2023 Release. Retrieved via the public Socrata API at data.cdc.gov (dataset id swc5-untb).
- [2] Appalachian Regional Commission (2026). County Economic Status and Distressed Areas by State, Fiscal Year 2027.
- [3] Appalachian Regional Commission (2026). About the Appalachian Region.
- [4] Meit, M., Heffernan, M., Tanenbaum, E., & Hoffmann, T. (NORC at the University of Chicago, for the Appalachian Regional Commission) (2017, updated through 2025). Appalachian Diseases of Despair.
- [5] Office of Disease Prevention and Health Promotion, U.S. Department of Health and Human Services. Social Determinants of Health, Healthy People 2030.
- [6] Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3), 351-357.